

The European Nucleotide Archive



Guy Cochrane

PhD University of East Anglia, 1999. At EMBL-EBI since 2002, Team Leader since 2009.

DESCRIPTION OF SERVICES

The European Nucleotide Archive (ENA) provides globally comprehensive primary data repositories for nucleotide sequencing information. ENA content spans the spectrum of data from raw sequence reads through assembly and alignment information to functional annotation of assembled sequences and genomes. Services for data providers include interactive and programmatic submission tools and curation support. Data consumers are offered a palette of services – sequence similarity search, text search, browsing, rich integration with data resources beyond ENA – all provided over the web and through an increasingly sophisticated programmatic interface. All ENA services are supported with a helpdesk and a growing training programme. These services are for users who approach ENA data and services directly, and those who provide secondary services (e.g. UniProt, Ensembl, Ensembl Genomes, ArrayExpress) that build on ENA content. Reflecting the centrality of nucleotide sequencing in the life sciences and the emerging importance of the technologies in applied areas such as healthcare, environmental and food sciences, ENA data and services form a core foundation upon which scientific understanding of biological systems has been assembled and our exploitation of these systems will develop. With ongoing focus on data presentation, integration within ENA, integration with resources external to ENA, tools provision and services development, the team's commitment is to the utility of ENA content and achieving the broadest reach of sequencing applications.

SUMMARY OF PROGRESS

- Enabled the capture, processing and presentation of major new raw read, read alignment, assembly and annotation data sets;
- Improved data submission pipelines for next-generation data from diverse platforms in a variety of formats;
- Further developed the ArrayExpress data-brokering scheme to allow metadata components of direct submissions from existing large-scale submitters (e.g. the Wellcome Trust Sanger Institute) to be routed to ArrayExpress curators;
- Deployed new similarity-search algorithms and datasets in ENA Sequence Search;
- Enabled community-informed development of a progressive compression strategy for sustainable ENA growth;
- Engineered and tested the CRAM Toolkit: a compression codebase;
- Significantly improved the ENA browser and text search, and delivered these improvements in a stepwise manner;
- Supported the ENA community by developing training courses and online training materials, and by delivering workshops.

MAJOR ACHIEVEMENTS

The ENA team launched nine new submission templates into our Webin system in 2011, bringing template coverage to the majority of curated submissions of assembled and annotated sequences. We also launched SRA Webin, a spreadsheet-based submission tool for interactive next generation sequence data submissions.

The development and launch of the ENA Taxon Portal and underlying data warehouse in 2011 has resulted in backend improvements and now provides richer real-time access to the massive data sets that we store.

We were key partners in a research project leading to the publication of a proof-of-principle paper on reference-based sequence compression. The method we proposed has

significant implications for the storage of raw sequence data, and accordingly for the reliability of the scientific record.

It is crucial that we are able to adapt quickly to changes in sequencing technology and to user requirements: accordingly, we lead a community-facing sequence read-compression initiative called CRAM. During 2011 we released the CRAM toolkit (in beta), a robustly engineered implementation of reference-based compression.

A major area of prototyping in 2011 was around the representation of assembly information. While ENA has traditionally captured and presented this information, the internal data structures and the public presentation of the content in these structures has been based upon concepts derived from flat file structure. As the complexity of assembly information grows, such as with the use of assembly patches

and alternative allele components, our traditional data structure has struggled to accommodate the storage of granular information of maximum utility to our consumers. Our prototype, a relational schema based on modelling of the assembly process itself and parallel work being carried out by our colleagues at the National Center for Biotechnology Information (NCBI), underwent testing with the UniProt, Ensembl and Ensembl Genomes teams and rounds of iterative improvement are in progress.

FUTURE PLANS

In 2012 we will focus on the consolidation of data submission services around the secure, dropbox-based model that has proven successful and flexible for SRA data submissions. Early in the year, we will deploy project registration services under this model and later development will be the addition of assembly submission services. In addition, we will prototype under our submissions services brokering to the EBI BioSamples Database.

Given the volume and dynamic nature of ENA content, a major challenge is the provision of data warehouses that can serve data to our interfaces in rapid and robust ways. In 2012 we will revisit our indexing and warehousing infrastructure and investigate the potential of new technologies to support rapid and frequent warehouse rebuilding (to support the latest content) and flexible interactive queries (to support such functionalities as faceted search and Boolean operations on search results).

In 2012, we expect our prototype for assembly information representation to reach maturity. We will then turn our attention to the ongoing curation of assembly information and the delivery of an easily consumed assembly data product.

Selected publications

Amid, C., Birney, E., Bower, L., *et al.* 2012. Major submissions tool developments at the European nucleotide archive. *Nucleic Acids Res.* 40, D43-7.

Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G., *et al.* 2011. Efficient storage of high-throughput DNA sequencing data using reference-based compression. *Genome Res.* 5, 734-40.

Karsch-Mizrachi I., Nakamura, Y., Cochrane, G. 2012. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40, D33-D37.

Kodama Y., Shumway, M., Leinonen, R., 2012. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40, D54-6.

We will continue to work on sustainable data archiving. In 2012, the CRAM toolkit is expected to mature from beta into production-ready code and we expect SRA submission pipelines, core data storage and presentation layers to use the CRAM format and toolkit to some extent. Our approach to CRAM toolkit development will continue to reflect a strategy that recognises the value of deep integration of reference-based compression in existing community tools and workflows but draws on the great potential that 'lossless' and 'lossy' compression offer.

In the context of CRAM, we will also focus on development and implementation of policy around the application of compression, specifically the nature of lossy compression models and the extent of data reduction applied under these models. Progressive compression, in which greater data reduction is applied to more reproducible sequence data (those data from highly available or reproducible samples) and greater data retention is applied to less reproducible sequence data (those data from rare and irreplaceable samples), is central to sustainable data archiving.

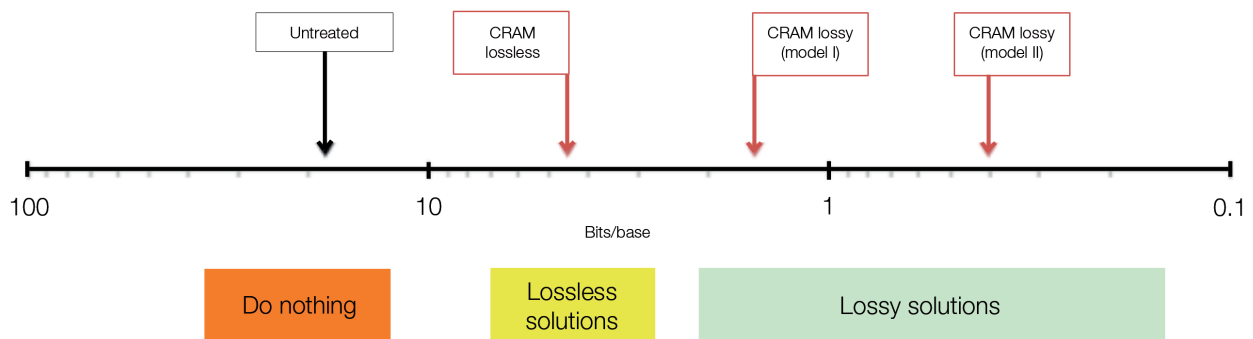


Figure. Reference-based compression scale: levels of compression in bits per base that can be achieved using CRAM reference-based compression. Current practice ('Untreated') is to use binary file formats, such as fastq and BAM. Lossless models under CRAM achieve around four-fold improvements in compression. Two lossy models are shown (models I and II) in which 10- and 50-fold compression can be achieved. Under more aggressive models, greater levels of compression can be achieved (not shown).